Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models

Francis K.C. Hui*[1], Samuel Müller†[2], and A.H. Welsh‡[1]

[1]Mathematical Sciences Institute, The Australian National University, Canberra, Australia

[2]School of Mathematics and Statistics, University of Sydney, Sydney, Australia

Abstract

In many applications of generalized linear mixed models (GLMMs), there is a hierarchical structure in the effects that needs to be taken into account when performing variable selection. A prime example of this is when fitting mixed models to longitudinal data, where it is usual for covariates to be included as only fixed effects or as composite (fixed and random) effects. In this article, we propose the first regularization method that can deal with large numbers of candidate GLMMs while preserving this

*Corresponding author: Francis K.C. Hui, Mathematical Sciences Institute, The Australian National University, 0200, Canberra, ACT, Australia. E: fhui28@gmail.com; P: +61 2 6125 0581.

†E: samuel.mueller@sydney.edu.au

‡E: alan.welsh@anu.edu.au

hierarchical structure: CREPE (Composite Random Effects PEnalty) for joint selection in mixed models. CREPE induces sparsity in a hierarchical manner, as the fixed effect for a covariate is shrunk to zero only if the corresponding random effect is or has already been shrunk to zero. In the setting where the number of fixed effects grow at a slower rate than the number of clusters, we show that CREPE is selection consistent for both fixed and random effects, and attains the oracle property. Simulations show that CREPE outperforms some currently available penalized methods for mixed models.

Keywords: fixed effects, generalized linear mixed models, LASSO, penalized likelihood, random effects, variable selection

1 Introduction

Joint selection of fixed and random effects in generalized linear mixed models (GLMMs) presents a challenging problem, especially as regards the question of how to perform selection in a computationally efficient manner while accounting for any hierarchical structure present in the model. Even with a bounded number of covariates, when jointly selecting over fixed and random effects the number of candidate models is considerably larger than in the standard regression context, making methods based on information criteria or the fence (Jiang et al. (2008)) computationally burdensome; see Müller et al. (2013) for a general review of model selection in linear mixed models. One approach to overcoming this computational problem is penalized likelihood methods. While penalized methods for generalized linear models have been extensively studied (dating back to Tibshirani (1996)), their application to mixed models has only recently been considered, almost exclusively in settings where the number of covariates is bounded, and the selection of fixed and random effects is treated as separate processes. Bondell et al. (2010) and Ibrahim et al. (2011) proposed separate penalties for the fixed and random effects that are summed together. Fan and Li

(2012), Peng and Lu (2012), and Lin et al. (2013) all proposed two-stage methods where the fixed and random effects selection are performed independently.

When fitting GLMMs to longitudinal data, there is a hierarchical structure in the selection of the effects that is often imposed in practice, namely "we usually only consider time-varying covariates that have been included in the fixed effects." (Cheng et al. (2010)). It is natural for covariates to be included as either a fixed effect only, or as both fixed and random effects. We refer to the latter as a *composite effect* covariate. As an example, in a longitudinal study monitoring the weights of infants over time (see Section 6), a random slope is included to account for heterogeneity between infants' changes in weight only if there is a significant overall trend (fixed effect) over time. Another example is in forest management, where random slopes are used to account for between plot variability only if a significant change is observed in the forest's overall health in response to climate (Hao et al. (2015)). Of course there may be exceptions to this hierarchical structure, a notable one being the case of linear mixed models with centered responses, where a random intercept may be included without a fixed intercept. For most settings however, it is reasonable that covariates should be included as either fixed or composite effects. However, while notions of hierarchical selection have been researched in (generalized) linear models with grouped variables and ordered or polynomial terms, see for instance the group LASSO (Least Absolute Shrinkage and Selection Operator) of Yuan and Lin (2006) and the composite absolute penalty of Zhao et al. (2009), they have not been investigated for GLMMs. This is exemplified in the illustrative examples of Bondell et al. (2010) and Ibrahim et al. (2011), where the respective penalties lead to at least one covariate selected only as a random effect.

We propose a penalty called CREPE (Composite Random Effects PEnalty) for hierarchical selection of fixed and random effects in longitudinal GLMMs. CREPE is the first penalty that directly incorporates the notion of covariates being selected as fixed or composite effects. This

is done by exploiting the hierarchical structure of the effects, such that a fixed effect coefficient is shrunk to zero only if the corresponding random effect coefficients are, or have already been shrunk to, zero. CREPE also accommodates covariates that are included *a-priori* as fixed effects only. The concept of using a penalty that accounts for the hierarchical structure of the effects has been considered in other contexts, e.g. the fused LASSO (Tibshirani et al. (2005)), finite mixture of regression models (Hui et al. (2015a)), and feature selection in bioinformatics (Garcia et al. (2014)), but has yet to be explored for joint selection in GLMMs. A key part of CREPE's design involves the use of a group-based penalty for selecting the random effects, specifically, the elements in a row of the eigendecomposition of the random effects covariance matrix (as defined in Section 2) are encouraged to be zero simultaneously.

In the setting where the number of fixed effects is allowed to grow at a slower rate than the number of clusters, we show that CREPE satisfies the oracle property of asymptotically identifying the truly non-zero fixed and composite covariates. Regarding computation, we use a Monte-Carlo Expectation Maximization (MCEM, Wei and Tanner (1990)) algorithm to calculate the CREPE estimates, showing how the E-step can be performed straightforwardly for the common cases of Gaussian, Poisson, and Bernoulli responses. Simulation studies show CREPE outperforms some other penalties available for jointly selecting fixed and random effects in GLMMs. We illustrate the application of CREPE to a longitudinal infant study for identifying important baseline and time-varying predictors of infant weights. We provide `R` code for calculating the CREPE estimates in the Supplementary Material; an `R` package is planned in future research.

2 Model Selection using CREPE

We focus on the independent cluster model with random intercepts and slopes. Let y_{ij} denote the j^{th} response collected for the i^{th} cluster, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. For simplicity, all clusters are assumed to have the same number of measurements, m, where m is bounded and does not grow with n. Conditional on the random effects, the y_{ij} are assumed to be independent responses from the exponential family $f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi)$ with mean μ_{ij} and dispersion parameter ϕ. Given a link function $g(\cdot)$, the mean is modeled as $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$ for a vector $\boldsymbol{x}_{ij}$ of predictors corresponding to fixed effects $\boldsymbol{\beta}$, and a vector $\boldsymbol{z}_{ij}$ of predictors corresponding to random effects $\boldsymbol{b}_i$, both containing an intercept term if appropriate. The random effects are assumed to have a multivariate Gaussian distribution, $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ and $\boldsymbol{\Gamma}$ is an unstructured matrix of the same dimension as $\boldsymbol{\Sigma}$, based on the eigendecomposition $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^T = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$ such that $\boldsymbol{\Gamma} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$, with $\boldsymbol{Q}$ an orthogonal matrix of normalized eigenvectors and $\boldsymbol{\Lambda}$ a diagonal matrix of eigenvalues.

Lemma 1. *Let $\boldsymbol{\gamma}_k$ be the k^{th} row of $\boldsymbol{\Gamma}$. Then for each k, $\|\boldsymbol{\gamma}_k\| = 0$ implies that $[\boldsymbol{\Sigma}]_{kl} = [\boldsymbol{\Sigma}]_{lk} = 0$ for all l, where $[\boldsymbol{\Sigma}]_{kl}$ refers to element (k, l) of $\boldsymbol{\Sigma}$, and $\|\cdot\|$ denotes the L_2-norm.*

This result suggests that, rather than penalizing the (diagonal) elements of $\boldsymbol{\Sigma}$ directly, we can employ a group-based penalty on the rows $\boldsymbol{\gamma}_k$, and indeed this is what we pursue. One advantage group-based penalization on the eigendecomposition has is that all the elements of $\boldsymbol{\Gamma}$ can take any number on the real line. This contrasts to the diagonal elements of both $\boldsymbol{\Sigma}$ and its Cholesky decomposition, which are bounded below by zero (see Bondell et al. (2010), Lin et al. (2013), and Pan and Huang (2014) for examples of methods that penalize the diagonal elements of $\boldsymbol{\Sigma}$ or its Cholesky decomposition). By using the eigendecomposition, we can avoid potential boundary issues when performing Taylor expansions (used in the theoretical study of the CREPE estimators in Section 3) and during the actual estimation

process.

For the independent cluster GLMM, the observed log-likelihood for a GLMM is,

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left(\int \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{\beta}, \phi, \boldsymbol{b}_i) f(\boldsymbol{b}_i|\boldsymbol{\Gamma}) d\boldsymbol{b}_i \right),$$

where $\ell_i(\boldsymbol{\Psi})$ is the log-likelihood contribution from the i^{th} cluster, and $\boldsymbol{\Psi} = \{\boldsymbol{\beta}, \phi, vec(\boldsymbol{\Gamma})\}$. We introduce some notation describing the nature of the covariates in the GLMM. Let α denote the full set of p covariates in the dataset. We divide this set into mutually exclusive subsets α_f, which denotes the set of p_f covariates entered into the model as fixed effects only (e.g., baseline covariates such as gender), and α_c, which denotes the set of p_c covariates entered into the model as composite effects (e.g., time varying covariates such as time of visit). We allow p_f to grow at a smaller rate than n (see Condition C6 in Section 3), while assuming $p_c < m$ is fixed. Subsequently, we can write $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \phi, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{p_c})$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\alpha_f}, \boldsymbol{\beta}_{\alpha_c})$.

The CREPE estimator is defined as the maximizer of the penalized log-likelihood function

$$\ell_{pen}(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - n\lambda \sum_{k=1}^{p} \tilde{w}_k \left(\beta_k^2 + \mathbb{1}_{\{k \in \alpha_c\}} \tilde{v}_k \|\boldsymbol{\gamma}_k\| \right)^{1/2}, \tag{1}$$

where $\lambda > 0$ is the tuning parameter and $\mathbb{1}_{\{.\}}$ denotes the indicator function. The adaptive weights $\tilde{w}_k$ and $\tilde{v}_k$ may depend on a common power parameter $\nu > 0$ (Zou (2006)) and are required to satisfy some regularity conditions.

For $k \in \alpha_f$, CREPE reduces to the adaptive LASSO penalty (Zou (2006)). On the other hand, for $k \in \alpha_c$, CREPE encourages sparsity in a hierarchical manner so that either both the fixed and random effects for the covariate are shrunk to zero, or only the random effect is shrunk to zero. There are two types of sparsity featured in CREPE: group sparsity, occurring on the rows of the eigendecomposition, $\|\boldsymbol{\gamma}_k\| = 0$, and the "larger" sparsity given

by $(\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_k\|\boldsymbol{\gamma}_k\|)^{1/2}$. Critically, the group sparsity is nested inside the larger sparsity event. Thus $\|\boldsymbol{\gamma}_k\| = 0$ must occur either before or simultaneously with $\beta_k = 0$. Then, in maximizing (1), CREPE allows a covariate $k \in \alpha_c$ to be included as either a fixed effect only, or as a composite effect.

Such a group penalty approach to random effects selection has been considered before by Ibrahim et al. (2011), and is arguably a better approach than that used by Bondell et al. (2010) amongst others, which penalizes the diagonal elements of the Cholesky decomposition of $\boldsymbol{\Sigma}$.

Fixed intercepts in GLMMs are generally not penalized, although the random intercept (if included) may be. In such a case, (1) can be altered to $\ell_{pen}(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - n\lambda(\tilde{v}_1\|\boldsymbol{\gamma}_1\|)^{1/2} - n\lambda\sum_{k=2}^{p}\tilde{w}_k(\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_k\|\boldsymbol{\gamma}_k\|)^{1/2}$, where it is assumed the first elements in $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ represent the fixed and random intercepts respectively.

3 Asymptotic Properties

We study the large sample properties of the CREPE estimator when p_f grows at a slower rate than n, while p_c is fixed. Allowing the number of random effects to grow is a more difficult problem, as it requires both the number of clusters and the cluster size to grow in order to achieve attractive asymptotic properties (see for instance Fan and Li (2012)), and (Demidenko (2004)) for an overview of asymptotic theory in mixed models.

Let $\boldsymbol{\Psi}_0 = (\boldsymbol{\beta}_0, \phi_0, \boldsymbol{\gamma}_{01}, \ldots, \boldsymbol{\gamma}_{0p_c})$, denote the true parameter values, where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0\alpha_f}, \boldsymbol{\beta}_{0\alpha_c})$ and, let p_{0f} be the number of non-zero elements in $\boldsymbol{\beta}_{0\alpha_f}$. Without loss of generality, we write $\boldsymbol{\Psi}_0 = (\boldsymbol{\Psi}_{01}, \boldsymbol{\Psi}_{02} = \mathbf{0})$ so $\boldsymbol{\Psi}_{01}$ consists of all the non-zero elements of $\boldsymbol{\beta}_0$, all the vectors $\boldsymbol{\gamma}_{0k}$ whose L_2-norm is positive, and ϕ_0. Likewise, we write the CREPE estimate as $\hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\Psi}}_1, \hat{\boldsymbol{\Psi}}_2)$. Let $H(\boldsymbol{\Psi}) = -(1/n)\partial^2\ell(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T$ denote the observed Fisher information matrix for the

GLMM, and let $\kappa_{\min}\{H(\boldsymbol{\Psi})\}$ and $\kappa_{\max}\{H(\boldsymbol{\Psi})\}$ denote its minimum and maximum eigenvalues respectively. The following regularity conditions are required here.

(C1) For every n, there exists a positive constant c_1 such that $0 < c_1 < \kappa_{\min}\{H(\boldsymbol{\Psi}_0)\} < \kappa_{\max}\{H(\boldsymbol{\Psi}_0)\} < 1/c_1 < \infty$.

(C2) For any given $\epsilon > 0$, there exists a $\delta > 0$ with $\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| < \delta$ such that $(1-\epsilon)c_1 < \kappa_{\min}\{H(\boldsymbol{\Psi})\} < \kappa_{\max}\{H(\boldsymbol{\Psi})\} < (1+\epsilon)/c_1$ for n large enough.

(C3) There exists an open subset Ω in the interior of the parameter space of $\boldsymbol{\Psi}$, containing $\boldsymbol{\Psi}_0$, such that the third derivatives of the log-likelihood $\ell(\boldsymbol{\Psi})$ exist for every $\boldsymbol{\Psi} \in \Omega$. For all $\boldsymbol{\Psi} \in \Omega$, there exist integrable functions U_{rst} such that $|\partial^3 \ell(\boldsymbol{\Psi})/\partial\Psi_r\partial\Psi_s\partial\Psi_t| < U_{rst}$, with $\mathrm{E}(U_{rst}^2) < \infty$, where the expectation is with respect to the true model.

(C4) $(\min_{l\in\boldsymbol{\Psi}_{01}}\{\beta_{0l}^2\} + \min_{l\in\boldsymbol{\Psi}_{01}}\{\|\boldsymbol{\gamma}_{0l}\|\}) \geq c_2$, where $c_2 > 0$ is a positive constant.

(C5) The adaptive weights satisfy $\tilde{w}_k = O_p(1)$ and $\tilde{v}_k = O_p(1)$ for $k \in \boldsymbol{\Psi}_{01}$, and $\tilde{w}_k = O_p\{(n/p_f)^{\nu/2}\}$ and $\tilde{v}_k = O_p\{(n/p_f)^{\nu/2}\}$ for $k \in \boldsymbol{\Psi}_{02}$.

(C6) (a) $\lambda\sqrt{np_{0f}} \to 0$ (b) $\lambda(n/p_f)^{(\nu+3)/4} \to \infty$, where $\nu > 0$.

Condition (C1) ensures the observed Fisher information matrix is well-defined at the true parameter values for every n, while condition (C2) extends this to a small neighborhood of $\boldsymbol{\Psi}_0$. The two conditions are similar to conditions A4 and A5 in Chen and Chen (2012) for generalized linear models (GLMs). Condition (C3) is a mild condition to ensure the log-likelihood function for GLMMs is sufficiently smooth. Since $\boldsymbol{\Psi}$ involves elements of the eigendecomposition $\boldsymbol{\Gamma}$ that can take any value on the real line, Ω is guaranteed to not lie on the boundary space. Condition (C4) places a lower bound on the magnitude of the truly non-zero coefficients. This may be weakened to permit the truly non-zero effects to tend

to zero at a slow rate, although we do not pursue this extension here. Together, conditions (C2) and (C4) define a rate at which incorrect models are allowed to approach the true model with increasing n. Condition (C5) is a generalization of condition (C1) in Ibrahim et al. (2011), requiring that the adaptive weights exhibit different asymptotic behavior for truly zero and non-zero coefficients. Finally, conditions (C6a) and (C6b) constrain the rate of growth of the tuning parameter λ, and is similar to conditions in Hui et al. (2015b) for adaptive LASSO GLMs. Together, they restrict the number of fixed effects to grow subject to $(p_f/n)^{(\nu+3)/4}\sqrt{np_{0f}} \to 0$. This is an advance on Ibrahim et al. (2011) and Lin et al. (2013), amongst others, who proved oracle properties assuming fixed p.

We first establish a result regarding the consistency properties of the CREPE estimator.

Theorem 1. *If (C1)-(C6) are satisfied and $\nu \geq 1$, then there exists a local maximizer $\hat{\boldsymbol{\Psi}}$ of the penalized log-likelihood function in (1) that satisfies*

(a) Estimation consistency: $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\| = O_p(\sqrt{p_f/n})$.

(b) Selection consistency: $P(\hat{\boldsymbol{\Psi}}_2 = \mathbf{0}) \to 1$.

With probability tending to one then, CREPE asymptotically correctly determines whether each covariate is a fixed or a composite effect.

Let $\mathcal{I}(\boldsymbol{\Psi}_0) = \mathrm{E}(-\partial^2\ell(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T)|_{\boldsymbol{\Psi}_0}$ be the expected Fisher information matrix evaluated at the true parameter point.

Theorem 2. *For a fixed integer q, let $\boldsymbol{B}_n$ be a $q \times \dim(\boldsymbol{\Psi}_{01})$ matrix such that $\boldsymbol{B}_n\boldsymbol{B}_n^T \to \boldsymbol{G}$ for some non-negative, symmetric $q \times q$ matrix $\boldsymbol{G}$. If (C1)-(C6) are satisfied and $\nu \geq 1$, then the local maximizer $\hat{\boldsymbol{\Psi}}$ in Theorem 1 satisfies*

$$\sqrt{n}\boldsymbol{B}_n\mathcal{I}^{-1/2}(\boldsymbol{\Psi}_{01})(\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{G}),$$

where $\mathcal{I}(\boldsymbol{\Psi}_{01})$ is the block of the expected Fisher information matrix involving only the truly non-zero parameters $\boldsymbol{\Psi}_{01}$.

Theorems 1 and 2 establish that the CREPE estimator attains the oracle property in GLMMs. The proofs of the theorems are provided in the Supplementary Material, following a similar outline to that of Fan and Peng (2004).

4 Estimation

We use the Monte-Carlo EM (MCEM, Wei and Tanner (1990)) algorithm combined with the local quadratic approximation (Fan and Li (2001)) for calculating the CREPE estimators. We focus on the common cases of Gaussian, Poisson, and Bernoulli mixed models, showing that updates of the parameters in these cases can be obtained straightforwardly. Let

$$\begin{aligned}\ell_{pen,c}(\boldsymbol{\Psi}, \boldsymbol{b}) &= \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\log\{f(y_{ij}|\boldsymbol{\beta}, \phi, \boldsymbol{b}_i)\} - \frac{1}{2}\log\{\det(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\} - \frac{1}{2}\boldsymbol{b}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\boldsymbol{b}_i\right) \\ &\quad - n\lambda\sum_{k=1}^{p}\rho(\beta_k, \boldsymbol{\gamma}_k) \\ &= \sum_{i=1}^{n}\ell_{c,i}(\boldsymbol{\Psi}, \boldsymbol{b}_i) - n\lambda\sum_{k=1}^{p}\rho(\beta_k, \boldsymbol{\gamma}_k)\end{aligned}$$

where $\rho(\beta_k, \boldsymbol{\gamma}_k) = \tilde{w}_k(\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_k\|\boldsymbol{\gamma}_k\|)^{1/2}$. Suppose at iteration t, we have estimates $\hat{\boldsymbol{\Psi}}^{(t)}$. The MCEM algorithm iterates between the following steps: the E-step, which calculates the expectation of $\ell_{pen,c}(\boldsymbol{\Psi}, \boldsymbol{b})$ with respect to the conditional posterior distribution $f(\boldsymbol{b}_i|\boldsymbol{y}, \hat{\boldsymbol{\Psi}}^{(t)})$, better known as the Q-function, and the M-step, which maximizes the Q-function to obtain updated estimates $\hat{\boldsymbol{\Psi}}^{(t+1)}$. For non-Gaussian responses where the posterior distribution does

not possess a closed form, we perform the E-step using Monte-Carlo integration,

$$\begin{aligned}\mathrm{E}_{\boldsymbol{b}_i|\hat{\boldsymbol{\Psi}}^{(t)}}\left\{\ell_{c,i}(\boldsymbol{\Psi},\boldsymbol{b}_i)\right\} &= \int \ell_{c,i}(\boldsymbol{\Psi},\boldsymbol{b}_i) \times \frac{\prod_{j=1}^{m} f(y_{ij}|\hat{\boldsymbol{\beta}}^{(t)},\hat{\phi}^{(t)},\boldsymbol{b}_i) f(\boldsymbol{b}_i|\hat{\boldsymbol{\Gamma}}^{(t)})}{\exp\{\ell_i(\hat{\boldsymbol{\Psi}}^{(t)})\}} d\boldsymbol{b}_i \\ &\approx \exp\{\ell_i(\hat{\boldsymbol{\Psi}}^{(t)})\}^{-1} \frac{1}{D}\sum_{d=1}^{D} \ell_{c,i}(\boldsymbol{\Psi},\boldsymbol{b}_i^d) \prod_{j=1}^{m} f(y_{ij}|\hat{\boldsymbol{\beta}}^{(t)},\hat{\phi}^{(t)},\boldsymbol{b}_i^d), \qquad (2)\end{aligned}$$

where $\boldsymbol{b}_i^d$ is simulated from $f(\boldsymbol{b}_i|\hat{\boldsymbol{\Gamma}}^{(t)})$, the quantity $\exp\{\ell_i(\hat{\boldsymbol{\Psi}}^{(t)})\}$ is approximated as $D^{-1}\sum_{d=1}^{D}\prod_{j=1}^{m} f(y_{ij}|\hat{\boldsymbol{\beta}}^{(t)},\hat{\phi}^{(t)},\boldsymbol{b}_i^d)$, and D is the number of Monte-Carlo samples. In the simulations in Section 5, we used $D = 2,000$.

To avoid non-differentiability at the origin, we approximate the CREPE penalty by a local quadratic approximation (LQA). At iteration t, set element k of $\hat{\boldsymbol{\Psi}}^{(t+1)}$ to zero if the corresponding element in $\hat{\boldsymbol{\Psi}}^{(t)}$ is equal to or very close to zero, e.g., absolute value within 10^{-3}. Otherwise, approximate the CREPE penalty as

$$\rho(\beta_k,\boldsymbol{\gamma}_k) = \rho(\hat{\beta}_k^{(t)},\hat{\boldsymbol{\gamma}}_k^{(t)}) + M_k^{(t)}(\beta_k^2 - (\hat{\beta}_k^{(t)})^2) + \mathbb{1}_{\{k\in\alpha_c\}} M_k^{(t)} \frac{\tilde{v}_k}{2\|\hat{\boldsymbol{\gamma}}_k^{(t)}\|}(\boldsymbol{\gamma}_k^T\boldsymbol{\gamma}_k - (\hat{\boldsymbol{\gamma}}_k^{(t)})^T\hat{\boldsymbol{\gamma}}_k^{(t)}),$$

where $M_k^{(t)} = (\tilde{w}_k/2)\left((\hat{\beta}_k^{(t)})^2 + \mathbb{1}_{\{k\in\alpha_c\}}\tilde{v}_k\|\hat{\boldsymbol{\gamma}}_k^{(t)}\|\right)^{-1/2}$. Combining these results, the M-step consists of maximizing the penalized Q-function,

$$Q_{pen}(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(t)}) = \mathrm{E}_{\boldsymbol{b}_i|\hat{\boldsymbol{\Psi}}^{(t)}}\left\{\ell_{c,i}(\boldsymbol{\Psi},\boldsymbol{b}_i)\right\} - n\lambda\sum_{k=1}^{p}\left(M_k^{(t)}\beta_k^2 + \mathbb{1}_{\{k\in\alpha_c\}} M_k^{(t)}\frac{\tilde{v}_k}{2\|\hat{\boldsymbol{\gamma}}_k^{(t)}\|}\boldsymbol{\gamma}_k^T\boldsymbol{\gamma}_k\right).$$

We now focus on the three special cases of Gaussian, Poisson, and Bernoulli responses.

Gaussian responses: For the linear mixed model where $f(y_{ij}|\boldsymbol{\beta},\phi,\boldsymbol{b}_i) = \mathcal{N}(\eta_{ij},\sigma^2)$, a closed form for the posterior distribution of $\boldsymbol{b}_i$ can be obtained. Let $\boldsymbol{y}_i = (y_{i1},\ldots,y_{im})$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}\ldots\boldsymbol{x}_{im})^T$ and $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}\ldots\boldsymbol{z}_{im})^T$. It is straightforward to show that $f(\boldsymbol{b}_i|\boldsymbol{y},\hat{\boldsymbol{\Psi}}) = \mathcal{N}(\hat{\boldsymbol{a}}_i,\hat{\boldsymbol{A}}_i)$, where $\hat{\boldsymbol{A}}_i = \left((\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^T)^{-1} + \hat{\sigma}^{-2}\boldsymbol{Z}_i^T\boldsymbol{Z}_i\right)^{-1}$ and $\hat{\boldsymbol{a}}_i = \hat{\sigma}^{-2}\hat{\boldsymbol{A}}_i\boldsymbol{Z}_i^T(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})$. In turn, we can derive a closed form for the penalized Q-function by using this result and the fact

that

$$\mathrm{E}_{\boldsymbol{b}_i|\hat{\boldsymbol{\Psi}}^{(t)}}(\boldsymbol{b}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\boldsymbol{b}_i) = \hat{\boldsymbol{a}}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\hat{\boldsymbol{a}}_i + \mathrm{tr}\{(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)^{-1}\hat{\boldsymbol{A}}_i\}, \tag{3}$$

an identity that does not require the normality assumption on $\boldsymbol{b}_i$. Closed form updates for $\boldsymbol{\beta}$ and σ^2 may then be obtained, while a Quasi-Newton method, for instance, can be used to update the rows of $\boldsymbol{\Gamma}$.

Poisson responses: Using the log link, we have $\sum_{i=1}^{n}\sum_{j=1}^{m}\log\{f(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i)\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\big\{y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta})\exp(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\big\}$. From this, it is straightforward to see that for the penalized Q-function, we only require Monte-Carlo estimates of the posterior mean $\mathrm{E}_{\boldsymbol{b}_i|\hat{\boldsymbol{\Psi}}^{(t)}}(\boldsymbol{b}_i)$, the moment generating function $\mathrm{E}_{\boldsymbol{b}_i}\{\exp(\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\}$, along with the posterior covariance matrix for use in (3). Since none of these is a function of the parameters that need updating, the M-step can be performed relatively quickly.

Bernoulli responses: Using the logit link, we have $\sum_{i=1}^{n}\sum_{j=1}^{m}\log\{f(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i)\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\big[y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \log\{1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\}\big]$. Applying the MCEM algorithm directly is challenging because the second term is non-linear in $\boldsymbol{\beta}$. To overcome this, we use the fact that the variance of the Bernoulli distribution is bounded above by 1/2. We can therefore minorize the above expression by a partial quadratic expansion about $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(t)}$,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}\log\{f(y_{ij}|\boldsymbol{\beta},\boldsymbol{b}_i)\} \geq \sum_{i=1}^{n}\sum_{j=1}^{m}\log\{f(y_{ij}|\hat{\boldsymbol{\beta}}^{(t)},\boldsymbol{b}_i)\} + \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\mu_{ij}^{(t)})\boldsymbol{x}_{ij}^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(t)}) \\ - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{m}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(t)})^T\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{(t)}),\end{aligned} \tag{4}$$

where $\eta_{ij}^{(t)} = \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{z}_i^T\boldsymbol{b}_i$ and $\mu_{ij}^{(t)} = \exp(\hat{\eta}_{ij}^{(t)})/\{1+\exp(\hat{\eta}_{ij}^{(t)})\}$ (see Hunter and Li (2005) for details on the notion of minorizing functions). Since this inequality remains true when we apply expectations to both sides, it means that we can use (4) to construct a minorizer

of $Q_{pen}(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(t)})$, and therefore maximize the minorizer instead. This is known as a (Monte-Carlo) minorization-maximization algorithm, as detailed in Hunter and Li (2005). Importantly, it is clear that this minorizer requires only Monte-Carlo estimates of $\mathrm{E}_{\boldsymbol{b}_i|\hat{\boldsymbol{\Psi}}^{(t)}}(\boldsymbol{b}_i)$, the expected fitted probability $\mathrm{E}_{\boldsymbol{b}_i}(\mu_{ij}^{(t)})$, along with the posterior covariance matrix for use in (3). As none of these is a function of the parameters that need updating, the maximization can be performed straightforwardly.

5 Simulation Study

An empirical study was conducted to compare the performance of CREPE with some other proposed penalties for variable selection in GLMMs. We focus on the cases of Gaussian, Poisson and Bernoulli responses. For brevity, only the results for Gaussian and Bernoulli mixed models are presented; the results for Poisson GLMMs are similar and are provided in the Supplementary Material. For CREPE, we chose the adaptive weights as follows. Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_f, \tilde{\boldsymbol{\beta}}_c)$ and $\tilde{\boldsymbol{\Sigma}}$ denote the maximum likelihood estimators of the fixed effects coefficients and random effects covariance matrix, based on fitting a saturated GLMM using the `lme4` package (Bates et al. (2014)). Then we set $\tilde{w}_k = |\tilde{\beta}_k|^{-2}$ and $\tilde{v}_k = [\tilde{\boldsymbol{\Sigma}}]_{kk}^{-2}$, where $[\tilde{\boldsymbol{\Sigma}}]_{kk}$ denotes the k^{th} diagonal element of $\tilde{\boldsymbol{\Sigma}}$. The saturated GLMM fit was also used to obtain starting values for the CREPE estimator. It is worth pointing out that the current version of `lme4` (version 1.1-10 at the time of writing) does not permit fitting mixed models when the number of random effects exceeds cluster size, $p_c > m$. Instead, we used an older version (version 1.0-6) that did permit such saturated models to be fitted.

In all three settings, we used a BIC-type criterion to select the tuning parameter for CREPE, $\mathrm{BIC}_\lambda = -2\ell(\hat{\boldsymbol{\Psi}}) + \log(n)\dim(\hat{\boldsymbol{\Psi}})$, where $\dim(\hat{\boldsymbol{\Psi}})$ denotes the number of *non-zero* estimated parameters in $\hat{\boldsymbol{\Psi}}$. The model complexity penalty used in the BIC is based on the log of the

number of clusters, n. More generally, our use of a BIC-type criterion for tuning parameter selection is comparable to what has been advocated in Bondell et al. (2010) and Lin et al. (2013), amongst others. We did however also consider the use of an AIC-type criterion, where $\log(n)$ was replaced by 2 as the model complexity penalty, with results (not shown) indicating that it tended to overfit both the fixed and random effects.

For each combination of n (number of clusters) and m (cluster size) considered, we generated 200 datasets. We assessed performance in terms of both model selection and model accuracy. For the former, we considered the mean number of false positives (truly zero coefficients not shrunk to zero, indicative of overfitting) and false negatives (truly non-zero coefficients shrunk to zero, indicative of underfitting) for the fixed effects, and the percentage of datasets with correctly chosen random effects. We also recorded the percentage of datasets where the method produced non-hierarchical shrinkage, where one or more covariates end up being selected as a random effect only. As discussed below (1), such non-hierarchical shrinkage is not permitted by the design of the CREPE penalty. In the Supplementary Material, we also present the percentage of datasets where the method obtained the correct model.

To assess model accuracy, we computed two measures for each method: the Kullback-Leibler distance between the true and fitted models, and the model error defined as the squared Euclidean norm between the estimated and true parameters. We subsequently computed a median relative Kullback-Leibler distance and the median relative model error, the median of the ratios of the Kullback-Leibler distance (or model error) between the CREPE estimator and the alternative method. Relative Kullback-Leibler distances and model errors less than one were indicative of CREPE having better model accuracy. Similar measures of model accuracy were used in Bondell et al. (2010) and Lin et al. (2013), among many others. Because the results for both measures were similar, we only present the relative Kullback-Leibler distance results in main text, and present the results for relative model errors in the

Supplementary Material.

5.1 Normal Responses

We adapted the simulation design in Bondell et al. (2010), but allowed the number of fixed effects to grow with n. In detail, datasets were simulated from a linear mixed model with the number of predictors growing at rate $p = \lceil 7n^{1/4} \rceil$ where $\lceil \cdot \rceil$ is the ceiling function. Covariates $\boldsymbol{x}_{ij}$ were constructed by setting the first element to one for a fixed intercept, and generating the remaining elements from a multivariate Gaussian distribution with mean zero and covariance matrix $\text{Cov}(x_{ijr}, x_{ijs}) = 0.5^{|r-s|}$. The covariates for the random effects $\boldsymbol{z}_{ij}$ were taken as the first eight covariates of $\boldsymbol{x}_{ij}$, so $p_c = 8$ and $p_f = p - p_c$ grows at the same rate as p. For the true model, the first eight elements of $\boldsymbol{\beta}_0$ were set to $(-1, 3, 1.5, 0, 0, 2, 1, 0)$. Then every third term was set to alternating values of ± 1. The true 8×8 covariance matrix for the random effects, $\boldsymbol{\Sigma}_0$, consisted of two non-zero blocks: I) a 2×2 matrix with diagonal entries 9 and 4, and off-diagonal entries of 4.8, occupying rows/columns 1 and 2 of $\boldsymbol{\Sigma}_0$, II) a 2×2 diagonal matrix with entries 2, occupying rows/columns 6 and 7 of $\boldsymbol{\Sigma}_0$. This resulted in four informative composite effect covariates. Responses y_{ij} were then generated from a Gaussian distribution with variance $\sigma_0^2 = 1$. We considered combinations of $n = 30, 60$ clusters (corresponding to $p = 17$ and 20 respectively) and cluster sizes of $m = 5, 10, 20$.

Three penalized estimators were compared: (1) CREPE with $\nu = 2$ in the adaptive weights for CREPE, (2) the M-ALASSO penalty of Bondell et al. (2010), and (3) the ALASSO penalty of Lin et al. (2013). To the best of our knowledge, these three procedures are currently the only penalties publicly available in `R` for selecting both fixed and random effects, and we found no additional methods. Since all procedures perform joint selection of fixed and random effects, we took the model error as $\text{ME} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 + \|\text{vech}(\hat{\boldsymbol{\Sigma}}) - \text{vech}(\boldsymbol{\Sigma}_0)\|^2$. Overall, CREPE performed the best in selecting both fixed and random effects, as well as

in model accuracy (Table 1 and Supplementary Material Table 1). M-ALASSO tended to choose a smaller number of fixed effects compared to CREPE, as reflected in the lower number of false positives but higher number of false negatives, while ALASSO performed worst as it severely overfitted the fixed effects. For random effects, M-ALASSO performed slightly better than CREPE although differences between the two were minor at the larger cluster sizes. For all settings, CREPE performed best in terms of selecting the correct model (Supplementary Material Table 1). ALASSO tended to underfit the random effects and shrink rows/columns 6 and 7 of the covariance matrix to zero. This underfitting of the random effects by ALASSO may be a result of the BIC used for the selecting the tuning parameter, which involves a large model complexity penalty $\log(mn)$ (following the recommendation in Lin et al., 2013). The median relative Kullback-Leibler distance was less than one in all but one case, indicating that CREPE has better model accuracy compared to the two alternative methods.

Both M-ALASSO and ALASSO presented cases of non-hierarchical shrinkage, particularly on element 7 in $\boldsymbol{x}_{ij}$ (and equivalently $\boldsymbol{z}_{ij}$) where the fixed effect was shrunk to zero while the corresponding random effect remained in the final model. Not surprisingly, the percentage of datasets where non-hierarchical shrinkage occurred decreased with increasing cluster size m.

5.2 Bernoulli Responses

We generated datasets from a Bernoulli GLMM using the same rate of growth of p (and thus p_f) as in Section 5.1. Covariates $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ were constructed in the same manner as in the Gaussian response case, $\boldsymbol{z}_{ij}$ being taken as the first eight covariates of $\boldsymbol{x}_{ij}$ such that $p_c = 8$. The elements of $\boldsymbol{\beta}_0$ were the same as in Setting 1, while the true 8×8 covariance matrix $\boldsymbol{\Sigma}_0$ was set to a diagonal matrix with the entries $(1, 1, 0, 0, 0, 1, 0, 0)$. Responses y_{ij} were then generated from a Bernoulli distribution with logit link. For CREPE, we used $\nu = 2$ for the

Table 1: Simulation results for linear mixed models. Performance was assessed the mean number false positives (FP) and false negatives (FN) for the fixed effects, the percentage of datasets with correctly chosen random effects components (%RE), percentage of datasets where there was non-hierarchical shrinkage (%S), and median relative Kullback-Leibler distance (RKL). Since %S was equal to zero for CREPE, this column is omitted from the table.

n	m	CREPE			M-ALASSO					ALASSO				
		FP	FN	%RE	FP	FN	%RE	%S	RKL	FP	FN	%RE	%S	RKL
	5	0.52	0.19	38	0.23	1.02	47	78	0.92	3.21	0.62	4	94	0.83
30	10	0.05	0.06	86	0.03	0.28	90	29	0.90	2.45	0.53	50	50	0.78
	20	0.06	0.02	95	0.01	0.24	96	24	0.50	4.46	0.42	41	35	0.39
	5	0.32	0.03	42	0.05	0.28	63	47	0.82	1.09	0.34	38	76	1.01
60	10	0	0.02	93	0	0.10	94	14	0.64	1.44	0.39	72	40	0.95
	20	0.01	0	97	0.01	0.07	96	9	0.49	3.37	0.31	56	39	0.63

adaptive LASSO weights. We considered combinations of $n = 50, 100$ clusters, corresponding to $p = 19$ and 23 respectively, and cluster sizes of $m = 10, 20$. We had intended to perform simulations at $m = 5$ also, as we did with Gaussian and Poisson responses, but found that we were unable to obtain suitable adaptive weights for CREPE based on a saturated GLMM fit. This was not surprising given the small cluster size $m = 5$ and relative lack of information in Bernoulli responses. While other methods of obtaining adaptive weights are possible, they are outside the scope of this work (see also our discussion in Section 7).

To our knowledge, no `R` packages are currently available for performing joint selection in mixed models with non-normal responses. For comparison with CREPE then, we considered the `glmmLasso` package (Groll and Tutz (2014)), which performs fixed effects selection only in GLMMs using the unweighted LASSO penalty. With this method, we considered two possibilities: the random effects component was known and only elements 1, 2, and 6 of $\boldsymbol{z}_{ij}$ were included; the random effects was unknown and all eight elements of $\boldsymbol{z}_{ij}$ were included. Our fitting models of such models via `glmmLasso` is unconventional in allowing fixed effects

to be penalized when the corresponding random effects (by definition of the program) cannot be penalized. We see this less as an argument against `glmmLasso` and more one in favour of using CREPE as a penalty.

Because `glmmLasso` only performs selection of the fixed effects here, the model error is based only on the fixed effects, $\text{ME} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$. This avoids confounding the results with whether the true and saturated random effects structure was used for `glmmLasso`. We considered several ways of implementing the package, and we present results based on the method which worked best, namely constructing a solution path from the smallest to the largest value of the tuning parameter.

CREPE performed better than both versions of `glmmLasso` at selecting the fixed effects, except at $n = 50$ and $m = 10$ where it had a slight tendency to underfit the fixed effects (Table 2 and Supplementary Material Table 3). This underfitting may explain why the relative Kullback-Leibler distance for both versions of `glmmLasso` was greater than one for this setting. In all other settings, CREPE had better model accuracy as reflected in the relative Kullback-Leibler distance (and model errors in Supplementary Material Table 2). At $n = 50$, both versions of `glmmLasso` tended to overfit the fixed effects, a result that may be partly attributed to the lack of adaptive weights. Regarding random effects selection, even at $n = 100$ and $m = 20$, CREPE was only able to correctly pick the true random effects structure half the time, with a tendency to overfit and fail to shrink rows/column 3 of the estimated $\boldsymbol{D}$ to zero (note this covariate has a corresponding non-zero fixed effect).

When the true random effects structure was known, `glmmLasso` presented no cases of non-hierarchical shrinkage (%S). By contrast, when a saturated structure was assumed for the random effects, strong evidence of non-hierarchical shrinkage was observed for `glmmLasso`, as it shrank one or more of the fixed effects for covariates 4, 5, and 8 to zero while leaving the corresponding random effects in the model. This was not surprising as our application

of `glmmLasso` allows fixed effects to be penalized in a situation where the program (by definition) cannot penalize the corresponding random effects.

Table 2: Simulation results for Bernoulli GLMMs. Performance was assessed based on the mean number false positives (FP) and false negatives (FN) for the fixed effects, the percentage of datasets with correctly chosen random effects components (%RE, for CREPE only), the percentage of datasets where there was non-hierarchical shrinkage (%S), and median relative Kullback-Leibler distance (RKL). Since %S was equal to zero for CREPE, the column is omitted from the table.

n	m	CREPE			$\texttt{glmmLasso}_{\text{true}}$				$\texttt{glmmLasso}_{\text{sat}}$			
		FP	FN	%RE	FP	FN	%S	RKL	FP	FN	%S	RKL
50	10	0.68	0.71	17	1.44	0.06	0	1.18	1.55	0.05	96	1.18
	20	0.13	0.01	31	2.54	0	0	0.74	3.55	0	87	0.70
100	10	0.15	0.02	11	0.57	0	0	0.85	0.78	0	100	0.82
	20	0.04	0	51	0.34	0	0	0.55	0.47	0	100	0.56

6 Application to Yale Infant Study

To illustrate the application of CREPE, we analyzed the Yale infant growth study of Wasserman and Leventhal (1993), which aimed to identify, among other things, whether cocaine exposure during pregnancy affects weight gain in children. The dataset was also used in Ibrahim et al. (2011). A total of $n = 298$ infants were recruited for the study, and their weight (in pounds) monitored over the study period. Seven predictors were available for analysis: gender of infant (1 for male; 0 for female), ethnicity (1 for African American; 0 otherwise), previous pregnancies (1 for yes; 0 for no), cocaine use by mother (1 for yes; 0 for no), age of mother (years), gestational age of infant (weeks), and day of visit during the study period (a proxy for time since entering the study). The number of visits for each infant ranged from $m = 2$ to $m = 30$, with a median of $m = 10$ visits. The goal of this analysis was

394 to identify important predictors of infant weight, while accounting for heterogeneity between
395 infants at baseline and over time.

396 It is natural to include the first four, time-independent covariates (gender, ethnicity, previous
397 pregnancies, cocaine use) in the model *a-priori* as fixed effects ($p_f = 4$), and to include the
398 three other time-varying covariates (age of mother, gestational age, day of visit) as composite
399 effects ($p_c = 3$). An intercept was also included in the model as a composite effect. Prior to
400 analysis, the three continuous covariates were standardized to have mean zero and variance
401 one. Adaptive weights were constructed by fitting the saturated model and setting $\nu = 2$.
402 Using BIC_λ to select the tuning parameter, the final model based on the CREPE estimator
403 had the following structure

$$\begin{aligned}\hat{\mu}_{ij} &= 6.962 - 0.190 \times \text{gender}_i + 0.245 \times \text{cocaine use}_i + 0.539 \times \text{gestational age}_{ij} \\ &\quad + 2.642 \times \text{visit}_{ij} + b_{0i} + b_i \times \text{visit}_{ij}; \\ \hat{\boldsymbol{D}} &= \begin{pmatrix} 0.548 & 0.277 \\ 0.277 & 0.214 \end{pmatrix}; \quad \hat{\sigma}^2 = 0.517.\end{aligned}$$

404 Of the four baseline covariates, CREPE identified gender and cocaine dependency as sig-
405 nificant predictors of infant weight. In particular, prenatal cocaine exposure (PCE) was
406 associated with higher infant weight, a surprising result given studies previously have found
407 significant evidence relating PCE and low birth weight (e.g. see the meta-analysis by Gouin
408 et al. (2011)). Of the time-varying covariates, CREPE identified gestational age as an impor-
409 tant fixed effect only, and day of visit as an important composite effect, with larger values of
410 both leading to higher overall infant weights. There was also significant variability between
411 infant weights at baseline as reflected in the inclusion of a random intercept, in addition to
412 the variability regarding how weights changed as a function of the day of visit.

413 Comparing the model chosen by CREPE to the one selected using the SCAD and IC_Q method

of Ibrahim et al. (2011) (see their Table 2), we find that the latter identified gestational age as (also) having an important random effect, and the age of the mother as having a significant random but not fixed effect, an example of non-hierarchical shrinkage. However, Ibrahim et al. (2011) did not include a random intercept as a candidate covariate, while in our analysis there was substantial variation between infants in their weights at baseline. It is of interest to point out that had we started with the saturated model and applied backwards elimination based on likelihood ratio tests (using `anova` with `lmer` in the `R` package), then this approach would have produced the same set of informative fixed and random effects as the model selected using CREPE.

7 Discussion

One avenue of research is to extend CREPE to ultra high-dimensional GLMMs, where the number of fixed and/or random effect potentially grows at a faster rate than the number of clusters and cluster size. Such an extension though is of more theoretical interest than of practical relevance. This extension is by no means straightforward: the adaptive weights require modification since the saturated GLMM can no longer be fitted using maximum likelihood estimation (e.g., weights might be constructed based on marginal models, Huang et al. (2008)), and the asymptotic theory demands growing n and m, differing assumptions on the degree of sparsity, and careful consideration of the differing impacts fixed and random effects have on the mixed model.

Supplementary Materials

The proof of Theorem 2, additional simulations results for Gaussian and Bernoulli GLMMs, full results for Poisson GLMMs, and `R` for implementing the CREPE penalty may be found in the Supplementary Material.

Acknowledgements

This research was supported by the Australian Research Council discovery project grant DP140101259. We are grateful to Andreas Groll for useful discussions.

References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-6.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–1077.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* **22**, 555–574.

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., and Muller, K. E. (2010). Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in Medicine* **29**, 504–520.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of statistics* **40**, 2043–2068.

Garcia, T. P., Müller, S., Carroll, R. J., and Walzem, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* **30**, 831–837.

Gouin, K., Murphy, K., and Shah, P. S. (2011). Effects of cocaine use during pregnancy on low birthweight and preterm birth: systematic review and metaanalyses. *American Journal of Obstetrics and Gynecology* **204**, 340.e1 – 340.e12.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing* **24**, 137–154.

Hao, X., Yujun, S., Xinjie, W., Jin, W., and Yao, F. (2015). Linear mixed-effects models to describe individual tree crown width for China-Fir in Fujian province, southeast China. *PloS one*, 10:e0122257.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Hui, F. K., Warton, D. I., and Foster, S. D. (2015a). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics* **9**, 866–882.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015b). Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association* **110**, 262–269.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33**, 1617–1642.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495–503.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics* **36**, 1669–1692.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics* **22**, 341–355.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science* **28**, 135–167.

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing* **24**, 725–738.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis* **109**, 109–129.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society* **67**, 91–108.

Wasserman, D. and Leventhal, J. (1993). Maltreatment of children born to cocaine-dependent mothers. *American Journal of Diseases of Children* **147**, 1324–1328.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* **68**, 49–67.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37**, 3468–3497.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.